Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Do You Mind Me Paying Less? Measuring Other-Regarding Preferences in the Market for Taxis

Brit Grosskopf, Graeme Pearce

To cite this article:
Brit Grosskopf, Graeme Pearce (2020) Do You Mind Me Paying Less? Measuring Other-Regarding Preferences in the Market for Taxis. Management Science

Published online in Articles in Advance 21 May 2020

. https://doi.org/10.1287/mnsc.2019.3483

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–16
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Do You Mind Me Paying Less? Measuring Other-Regarding Preferences in the Market for Taxis

Brit Grosskopf,[a] **Graeme Pearce**[a]

[a] Department of Economics, University of Exeter Business School, Exeter, Devon EX4 4PU, United Kingdom
Contact: b.grosskopf@exeter.ac.uk, https://orcid.org/0000-0002-6535-5676 (BG); g.pearce@exeter.ac.uk, https://orcid.org/0000-0002-4701-2629 (GP)

Received: October 15, 2017
Revised: May 14, 2018; March 1, 2019; July 5, 2019
Accepted: July 26, 2019
Published Online in Articles in Advance: May 21, 2020

https://doi.org/10.1287/mnsc.2019.3483

Abstract. We present a natural field experiment designed to measure other-regarding preferences in the market for taxis. We employed testers of varying ethnicity to take a number of predetermined taxi journeys. In each case, we endowed them with only 80% of the expected fare. Testers revealed the amount they could afford to pay to the driver midjourney and asked for a portion of the journey for free. In a 2 × 2 between-subjects design, we vary the length of the journey and whether a business card is elicited. We find that (1) the majority of drivers give at least part of the journey for free, (2) giving is proportional to the length of the journey, and (3) 27% of drivers complete the journey. Evidence of outgroup negativity against black testers is also reported. In order to link our empirical analysis to behavioral theory, we estimate the parameters of a number of utility functions. The data and the structural analysis lend support to the quantitative predictions of experiments that measure other-regarding preferences, and they shed further light on how discrimination can manifest itself within our preferences.

History: Accepted by Yan Chen, behavioral economics.

Funding: The authors thank the University of Exeter Business School for funding this research.
Supplemental Material: Data and the online appendix are available at https://doi.org/10.1287/mnsc.2019.3483.

Keywords: other-regarding preferences • field experiment • discrimination

1. Introduction

A large number of experiments detail the prevalence and significance of other-regarding preferences. Other-regarding preferences incorporate the idea that the traditional assumption of self-interest must sometimes be relaxed to account for interdependent preferences, such as fairness and envy. This class of preferences relates to the consequence of outcomes and assumes that individuals have a preference over how payoffs are distributed between themselves and others. They have been found to be important for behavior in a range of laboratory interactions, including social dilemma games, dictator games, and coordination games (Camerer and Fehr 2004, Cooper and Kagel 2009).

However, evidence from the field is mixed. For example, Stoop et al. (2012) and Winking and Mizer (2013) find no evidence of other-regarding preferences in public goods and dictator games conducted in the field, whereas Stoop (2014) does. Others have found that monitoring considerations explain behavior consistent with other-regarding preferences (Bandiera et al. 2005, Benz and Meier 2008), and List (2006) highlights the importance of reputational concerns. By contrast, there exists a large literature that has found evidence of other-regarding preferences (Landry et al. 2006, DellaVigna et al. 2012) in the market for charitable giving. There are, however, relatively few studies that have examined the implications of other-regarding preferences for behavior in what might be considered more "traditional" markets (i.e., those that do not rely on discretionary gifts but on the exchange of goods and services).

The purpose of this paper is to examine the extent to which other-regarding preferences can shape the behavior of economic agents operating in a marketplace. It is important to investigate this in the field because theoretically, as shown by Dufwenberg et al. (2011) in a general equilibrium model, agents assumed to have other-regarding preferences could behave as if they were selfish. We also investigate the significance of ethnic identity in determining these preferences because other-regarding preferences form the foundation for recent behavioral theories of discrimination. Stemming from concepts of "taste-based" discrimination first detailed by Becker (1971), a prominent theory is that other-regarding preferences are group

contingent, or larger toward those we identify with (the "ingroup") in comparison with those we do not ("outgroups"; Chen and Li 2009). Although this explanation has gained prominence, as with other work on social preferences, the majority of evidence in its support has been obtained from laboratory experiments (Goette et al. 2006, van Der Mewe and Burns 2008, Chen and Chen 2011, Drouvelis and Nosenzo 2013).[1] Field experiments, by contrast, largely suggest that discriminatory behavior can be attributed to statistical discrimination (Levitt 2004, List 2004, Gneezy et al. 2015), although it is not always possible to distinguish between the two explanations (Bertrand and Mullainathan 2004, Mujcic and Frijters 2013).

The study closest to ours is that of Mujcic and Frijters (2013). Exploiting a natural interaction between bus drivers and passengers, they paid testers to act as passengers to try to board buses without any money. They find that white testers are allowed to embark 72% of the time, Indian testers 51%, and black testers just 36%. The interaction can be viewed as an other–other allocation game (Tajfel et al. 1971), where the driver must allocate resources between the passenger and the bus company, rather than being comparable to the dictator game. Because drivers are not monitored, their choices, although costly to the bus company, are financially costless to themselves. Our study distinguishes itself from Mujcic and Frijters (2013) in a number of important and economically significant ways. First, we consider discrimination in a situation where prosocial behavior is financially costly to the person exhibiting it. Unlike the employed bus drivers studied by Mujcic and Frijters (2013), we examine self-employed taxi drivers who bear all the financial costs of their decisions. Second, we investigate an interaction that is not scrutinized by bystanders, as is the case in Mujcic and Frijters (2013), where other passengers observe the bus driver's decision.

We conduct a field experiment to examine the behavior of taxi drivers in Great Britain. We employed 22 testers of varying ethnicity to pose as passengers and take a number of predetermined taxi journeys.[2] In each case, we endowed them with only 80% of the expected fare. Once the taxi meter reached 60% of the fare, testers told the driver that they only had a certain amount and asked if they could have the final 20% of the journey for free. The tradeoff faced by a driver in this transaction is analogous to the dilemmas that subjects typically face in the laboratory: express other-regard at a personal cost but to the benefit of another by giving some of the journey for free or behave selfishly but profitably by stopping once the meter reaches the amount the passenger can afford.[3]

In a 2×2 between-subjects design, we systematically vary the length of the taxi journeys using short- and long-distance treatments, where testers take journeys of approximately 1.7 miles and 4.4 miles. Because drivers assigned to the long-distance treatment are able to give twice as much (in absolute terms) as drivers assigned to the short-distance treatment, we can examine whether the drivers' other-regarding preferences depend on the relative payoffs between themselves and the passenger or if giving is constant regardless of the amount available to give. The taxi markets we study have thousands of drivers and tens of thousands of passengers each week, making repeated interactions for infrequent customers incredibly unlikely. These markets are therefore attractive for studying the "one-shot" interactions required for disentangling other-regard from reputational concerns, and the only real possibility of meeting a driver in a future interaction is by obtaining his or her contact details so that he or she can be actively selected. In the baseline treatments, testers reinforce the one-shot nature of the interaction by stating to the driver that they "don't take taxis very often." However, as shown by List (2006), field experiments designed to detect social preferences need to be particularly careful about the possibility of reputational concerns. To address this, and to reinforce our interpretation of the one-shot nature of the baseline treatment, we conduct a business card treatment to examine whether drivers are willing to give out their contact details for potential future interactions.

We find that 74% of drivers in the baseline treatment give part of the journey for free, with 27% completing the journey at no extra cost to the tester. We also find that the extent of giving is proportional to the length of the journey, with drivers giving about 10% of the expected fare in both short- and long-distance treatments. In addition, drivers do not seem concerned about repeated business with customers, with only 45% producing a business card when asked to do so. Drivers who fail to give a business card behave identically to those assigned to the baseline treatment. This demonstrates the inherent one-shot nature of the interactions in the market we study, and we feel confident that the drivers' behavior from the baseline treatments is not influenced by reputational concerns. However, drivers assigned to the business card treatment who do provide a business card are found to give significantly more than those who do not, but only for short-distance journeys.

Differential treatment of testers conditional on both their own and the drivers' ethnicity is also observed: white and South Asian drivers give significantly less, and they are significantly less likely to complete a journey when the tester is black. This result is robust to a comprehensive range of field-, journey-, driver-, and tester-specific variations obtained from each individual journey. Tester-specific characteristics are obtained from a complementary laboratory experiment, following the

procedure of Xiao and Houser (2005). We elicit the perceived aggressiveness, attractiveness, friendliness, trustworthiness, and wealthiness of the testers' appearance, traits that are otherwise "unobservable" but may vary with ethnicity (Heckman 1998). To link our results to behavioral theory, we also conduct a structural analysis in order to obtain other-regarding preference parameter estimates. Estimates from a range of models reveal that the other-regarding preferences of drivers are qualitatively and quantitatively similar to those obtained from laboratory experiments and that these preferences are group contingent. Although our results are consistent with the theory of group-contingent social preferences (i.e., taste-based discrimination), we cannot rule out statistical explanations (Phelps 1972).

This study makes a number of contributions. First, we provide evidence that other-regarding preferences can shape the behavior of market agents with a similar prominence to subjects in the laboratory. Second, we demonstrate that the observed behavior cannot be attributed to drivers' reputational concerns. Finally, we show that a model that assumes other-regarding preferences to be group contingent is able to explain the observed behavior.

The remainder of this paper is organized as follows. Section 2 discusses the taxi markets we study, and Section 3 details the experimental design. Section 4 outlines reduced-form estimation results, and it discusses the estimates from the structural model. Section 5 discusses alternative interpretations of the results, and Section 6 concludes.

2. The Market for Taxi Services

In Great Britain, there are two types of vehicles that operate as taxis: private hire vehicles (PHVs) and Hackney carriages. PHVs are not as strictly regulated as the latter, and anyone who has a driving license and is willing to pay the licensing fee, in practice, is able to become a PHV driver. PHVs are unable to ply for hire and must be prebooked over the phone: passengers must actively select a company or driver for a given journey. The price of the journey (or fare) is independently set by each firm or negotiated ex ante, and vehicles often do not have a fitted meter. As such, PHV fares can vary wildly, as can the types of vehicles used.

By contrast, Hackney carriages are taxis in the true sense: drivers can ply for hire, with customers able to hail or call them, and drivers able to wait at designated taxi ranks to be approached by customers. Drivers and passengers are randomly matched, and importantly, customers are unable to select their driver. When hailing a vehicle, a customer must take whichever driver happens to be in the area. At a rank, customers must take the taxi at the front of the queue, and drivers further down the queue will refuse journeys from customers who approach them. The only real possibility of using the same driver repeatedly is by obtaining his or her personal contact details.

The strict regulation of Hackney carriages ensures their similarity, with all drivers having to pass a road knowledge and English language test. All vehicles have to adhere to strict standards, such as being fitted with safety screens to separate the driver and passenger, having wheelchair access, and having the vehicle being under a certain age.[4] All vehicles are fitted with a taxi meter that displays the cost of the journey, up to a given point, to the passenger. The meter starts from a fixed amount and increases by a set amount every so many yards driven or seconds waiting in traffic. Metered fares are set by the local authority. Those relevant for this study are detailed in Table A1 in Online Appendix A.[5]

Important to our study is the fact that the metered fare is the maximum fare the driver is able to charge the passenger. Fare reductions are made entirely at the driver's discretion, and the driver is within his or her rights to refuse any reductions the passenger asks for. The 2014 Birmingham Unmet Taxi Demand Survey indicates that the vast majority of Hackney carriages (90%) are driver owned: drivers keep all the fare and tips,[6] and incur all the costs associated with a journey.[7] The cost of a discretionary fare reduction is therefore borne exclusively by the driver.

The markets we study are incredibly thick, with tens of thousands of journeys taken each week and with over a thousand licensed Hackney carriages operating in each city. As outlined in Table A3 in Online Appendix A, some of the taxi ranks see more than 19,000 passengers per week. The vast majority of taxi journeys are taken from taxi ranks, and at a taxi rank, a driver has to wait, on average, just 12 minutes for his or her next passenger.[8] The sheer number of transactions, large number of taxi ranks, and ability of drivers to "cruise" streets plying for hire mean that an infrequent user of Hackney carriages is highly unlikely to have a repeated interaction with the same driver, and the driver the user does interact with is essentially randomly assigned.

3. Experimental Design and Procedure

The experiment was designed to measure other-regarding preferences of Hackney carriage drivers (hereafter referred to as "taxi drivers") in actual market transactions and determine the extent to which these preferences vary with their own and the passenger's ethnicity. We use a field experiment that allows us to observe behavior in a market setting, in a natural interaction devoid of experimenter scrutiny. Our subjects, the taxi drivers, were oblivious to a study taking place.

3.1. Testers

The testers were hired by placing a job advert looking for "research assistants" on the Universal Jobsmatch website, a national website initiated by the UK government's Department for Work and Pensions, which anyone can use to advertise a job. The advert stated that individuals were required to assist in conducting some "economic research." Although the specific job role was not stated, it was advertised that some walking in and around the city center would be required. Everyone who applied was invited to attend a briefing and training session at a neutral location, where they were told about the job role and asked to sign consent forms in order to take part. The rate of pay was £8.30 per hour (all experimental materials are given in Online Appendix A).[9] We hired 22 testers in total. This compares favorably with previous studies of taxi markets that have typically employed just a handful of testers.[10]

Briefing sessions lasted between one and two hours, and a single treatment was discussed in detail. Testers were given copies of *one* script they were required to follow and the experimental sheet they would have to complete.[11] They were told that the script may vary and that they would be given a chance to practice any variants before completing the task. Testers were told explicitly to follow the script as closely as possible, and when interacting with the drivers, they were told they must not attempt to influence any of their decisions. Testers were told not to engage in conversation with the drivers, and scripted responses were given to anticipated questions. Our hypotheses and predictions regarding the study were never made clear to the testers, and not all the testers met each other, reducing the opportunity for testers to guess that the study might involve their own ethnicity.[12] All testers wore casual clothing.[13]

Each tester also consented to have his or her face photographed for "research purposes." Once the field experiment was complete, we had the testers' appearance rated by subjects in a follow-up laboratory experiment. Subjects in the laboratory had to rate the pictures for aggressiveness, attractiveness, friendliness, trustworthiness, and wealthiness on a scale from 1 to 10 (with 1 being "not very" and 10 being "very"). This was done to control for otherwise unobservable characteristics that may vary with the testers' ethnicity (Heckman 1998), although we acknowledge that they are entirely subjective.

These five characteristics were chosen for a number of reasons. First, the importance of an individual's attractiveness in fostering the helping behaviors of others has been outlined in a wealth of studies, with the most attractive typically found to be treated most generously (Benson et al. 1976). Attractiveness has also been shown to be successful in promoting others' other-regarding behaviors (Landry et al. 2006) and is correlated with labor market outcomes (Mobius and Rosenblat 2006). Second, historical and recent evidence suggests that faces that appear aggressive and unfriendly, or threatening, may stimulate a different thought system in comparison with one seen as nonthreatening. For example, Öhman (1986) argues that threatening faces activate the "fear system" and therefore provide a powerful stimulus. If this is the case, faces displaying differing levels of aggression and friendliness may trigger different types of behaviors, such as self-defensive compared with helping behaviors (see Schupp et al. (2004) for evidence and a discussion of the literature). Third, any differential in giving stemming from ethnicity may be related to status differences relating to wealth, similar to that shown by Mitra and Ray (2014). Finally, as the interaction between a driver and tester may rely on the driver trusting the passenger regarding how much money he or she has, we also elicit the passengers' facial appearance of trustworthiness.

To obtain the ratings, we follow a procedure similar to that used by Landry et al. (2006). Each laboratory subject was shown a random set of 11 photos and asked to rate their appearance. Following Xiao and Houser (2005), to increase subjects' attentiveness to the task, they were told that one photo and one characteristic of that photo would be selected at random, and if their decision for that photo and that characteristic was in line with the ratings of the majority of the other subjects in the session, they would receive £2. It took subjects about 10 minutes to rate all the photos required of them.[14] A sample of 1,188 ratings of the 22 testers was obtained from 108 laboratory subjects. The testers' characteristics, along with these ratings, are presented in Table B6 in Online Appendix B.[15]

We find that black testers are rated significantly less attractive, trustworthy, friendly, and wealthy than both white and South Asian testers ($p < 0.001$ in all cases, robust rank-order tests). Black testers are also rated the most aggressive ($p < 0.001$ in both cases, robust rank-order tests). Interestingly, white testers are rated as less attractive, trustworthy, friendly, and wealthy than South Asian testers ($p = 0.06$ for attractiveness, $p < 0.001$ in all other cases, robust rank-order tests). White testers are also seen as more aggressive than South Asian testers ($p < 0.001$, robust rank-order test). We control for these tester-specific variations in our parametric analysis in Section 4.

We focus exclusively on facial appearance because of the way that the driver and tester interact while in the taxi. As outlined in Section 3.2, the driver's decision to behave in an other-regarding manner is made while he or she is driving, so he or she is likely to view the tester briefly, either through the rearview mirror or by looking over his or her shoulder. Visual emphasis will be placed on the tester's face rather

than on other physical traits, such as body mass index, height, or build.

3.2. Procedure

On a given day, a tester was required to complete multiple journeys ranging from 3 to 10. Journeys (designated as short and long) were randomly assigned to testers. Because the journeys were taken from ranks, the testers had to wait for their turn to approach the taxi at the front of the rank, enter the taxi, and then state their destination.[16] The experiment first varies the distance of the journeys in short- and long-distance treatments, with respective journey lengths of approximately 1.7 miles and 4.4 miles, which had expected fares of approximately £5 and £10, respectively. The journeys were calibrated using Google Maps and the relevant taxi fare tables. The journeys always started from a rank,[17] and destinations were landmarks and well-known locations that did not have a designated taxi rank.[18] The testers were endowed with either £4 or £8 for each journey, depending on its distance. The journeys were taken in either Birmingham or the greater Manchester area, with those starting in Birmingham taken over five days and those in Manchester over three. All journeys were taken between 11 a.m. and 5 p.m., and at least four testers were in the field at any given time, along with an experimenter. Testers took part in multiple treatments across different days.

On entering the taxi, the tester first stated his or her destination and then spoke a simple entry statement.[19] In the baseline treatment, the tester stated, "I don't take taxis very often"; in the business card treatment, he or she stated, "I'm looking for a reliable driver for future journeys. Can I have a business card?" The first statement signals to the driver that the interaction is one shot, because a passenger who does not take taxis very often is unlikely to meet the same driver twice. The second statement was chosen following the repeat business treatment of a field experiment conducted by Schneider (2012). The statement gave the driver the opportunity to give out a business card before the journey began. The proportion of cards given out should provide some indication of the drivers' concern for repeated business, with a higher proportion signaling a higher concern. The scripts were designed to be kept simple in order not only to keep them standardized and to avoid actor bias (Heckman 1998) but also to keep them natural and believable to the drivers. This design feature clearly contrasts with laboratory experiments, where interactions are designed to be "sterile" and, predominantly, without context.[20]

Once the taxi journey began, the testers were required to wait in silence until the meter reached a certain amount: £3 in short- and £6 in long-distance journeys, or 60% of the expected fare. Once the meter reached this amount, testers stated to the driver, "I'm sorry, I only have £x! Can you still take me to my destination for that amount?," where x = £4 in short- and x = £8 in long-distance journeys. By revealing this to the driver once the meter reached 60% of the expected fare, the driver was given ample time to stop the taxi. It also signaled the testers' intention to pay the amount that they could afford, removing any belief the driver may have that the passenger would not pay. Table 1 summarizes the experimental design.

We refer to the driver continuing the journey past the amount that the tester can afford as *giving*, which is accurately measured by the meter. We define giving in this way because it provides a reasonable lower bound of the drivers' opportunity cost. This is because a driver can always return to a rank and pick up his or her next passenger. A follow-up survey conducted at taxi ranks in Manchester, presented in full in Section 5, reveals that the vast majority of taxi journeys begin at a taxi rank and that the majority of taxi drivers return to the rank from which the journey started. Furthermore, the 2015 Unmet Taxi Demand Survey published by the Manchester City Council suggests that the average wait time for a passenger at a taxi rank is very short (12 minutes). Regardless of how the driver works, on fixed shifts or income targeting (Camerer et al. 1997), the time spent driving the passenger beyond the amount he or she can afford is time spent not earning.

Once the driver decided how much to give and where to end the journey, the tester had to ask for a receipt, leave the taxi, and discreetly complete an experimental sheet. The sheet included subjective characteristics of the driver, such as his or her age, gender (1 if male), and ethnicity; measures of the field, including traffic intensity (recorded on a 10-point scale: 1 if not busy, 10 if very busy) and the weather (1 if raining); and finally, characteristics of the ride, including whether the driver attempted a conversation (1 if yes), if he or she offered to drive to a cashpoint/automatic teller machine (ATM; 1 if yes), and (in the business card treatment) if he or she gave a business card or not (1 if one was given). Most important, the testers had to record the final meter reading

Table 1. Experimental Design Summary

		Short distance	*Long distance*
Baseline	Entry script	"I don't take taxis very often."	
	Endowment (£)	4	8
	Expected fare (£)	5	10
Business card	Entry script	"I'm looking for a reliable driver for future journeys. Can I have a business card?"	
	Endowment (£)	4	8
	Expected fare (£)	5	10

Note. The expected fare of journeys in each treatment is approximate.

and whether the driver completed the journey or not.[21] Table 2 presents the average characteristics associated with the driver,[22] the field conditions, and the ride, as reported by the testers.[23]

4. Results

In this section, we outline the experimental results. A number of common features are present throughout the analysis. Where nonparametric tests are used, both the p-value and the test statistic are presented in parentheses. Unless otherwise stated, all tests are two-sided, and journeys from all treatments are pooled in the regressions.

4.1. Journey Calibration Checks

Some initial calibration checks are conducted in order to examine whether our expected fare calculations are accurate. Table 3 outlines the recorded fare, expected fare, and amounts given as a percentage of the expected fare from journeys where the driver completed the journey. Observations are disaggregated by short- and long-distance journeys. By comparing the observed fare of a completed journey with its expected fare, we can examine the accuracy of our expected fare calculations. This will also shed light on how the drivers perceived the testers (i.e., as locals or nonlocals; Balafoutas et al. 2013). Minor discrepancies between recorded and expected fares are to be expected largely because of variations in traffic intensity and other random shocks.

Table 2. Driver, Field, and Ride Characteristics

Driver characteristics		Driver ethnicity			
	All drivers	White	Black	South Asian	Other
Age	44.34 (10.67)	50.06 (10.56)	40.36 (9.36)	42.60 (10.03)	41.33 (11.45)
Gender (1 if male)	0.99 (0.12)	0.97 (0.17)	1.00 (0.00)	1.00 (0.10)	1.00 (0.00)
Journeys	283.00	71.00	11.00	191.00	10.00
Field characteristics					
Traffic Intensity (1 = not busy; 10 = very busy)					4.44 (2.26)
Weather (1 if raining; 0 otherwise)					0.11 (0.32)
Ride characteristics					
Conversation (1 if driver attempted a conversation)					0.28 (0.45)
Cashpoint (1 if driver offered a cashpoint)					0.04 (0.20)
Business card, business card treatment only (1 if given)					0.45 (0.50)
Receipt (1 if given)					0.89 (0.31)

Notes. Standard deviations are in parentheses. Where the driver's ethnicity is classified as "Other," the tester either did not complete the experimental sheet or classified the driver outside the three main ethnic groups that are specified.

Table 3. Fares, Expected Fares, and Average Giving Conditional on the Driver Completing the Journey

Parameter	*Short distance*	*Long distance*
Recorded Fare (£)	5.44 (1.30)	10.43 (1.47)
Expected Fare (£)	5.65 (0.22)	10.22 (0.93)
Amount Given, % Exp. Fare	26.00 (0.23)	24.00 (0.15)
Completed journeys	44	22

Notes. We exclude from these calculations 18 observations where the driver completed the journey but switched off the meter before the journey was completed. In these 18 cases, in Section 4.2, we approximate the meter reading by the expected fare. Standard deviations are in parentheses.

Formally comparing the recorded and expected fares, we report no significant differences in the short-distance treatment ($p = 0.304$, t-test) or long-distance treatment ($p = 0.539$, t-test). The experiment was designed so that the driver would have to give about 20% of the expected fare for free in order to complete the journey. Examining this, we find that the amount given as a percentage of the expected fare in the observed completed journeys is not significantly different from the planned 20% in both the short- ($p = 0.88$, sign test) and long-distance ($p = 1$, sign test) treatments. Therefore, we conclude that the journey planning is accurate.

4.2. Other-Regard

Table 4 outlines average amounts given by drivers and the proportion of journeys they completed by treatment and by the testers' ethnicity. To examine whether relative payoffs are a motivating factor behind the amounts that drivers are giving, we report giving as a percentage of the expected fare. Figure 1 displays the distribution of giving from each treatment.

Table 5 reports a number of random effects Tobit regressions. In models (1)–(4), giving in pounds by driver i to tester j is the dependent variable. In models (5)–(12), giving as a percentage of the expected fare by driver i to tester j is the dependent variable. Considering giving in this way enables us to control for the variation in journey lengths and therefore variation in the expected fares of journeys, both within and between treatments. In each regression, dummy variables for the long-distance treatment and the business card treatment ($BusC$) are included along with their interaction; the short-distance baseline treatment is taken as the control. Because receiving a business card from the driver, rather than simply asking for one, is what most likely activates any repeat

Table 4. Average Driver Giving by Treatment and Testers' Ethnicity

		Baseline		Business Card	
Testers' ethnicity	Parameter	Short	Long	Short	Long
White	*Amount Given* (£)	0.64	1.23	0.87	1.22
		(0.62)	(1.44)	(0.74)	(1.31)
	Amount Given, % Exp. Fare	11	13	16	12
		(0.11)	(0.16)	(0.13)	(0.13)
	Number of Journeys	60	26	49	29
Black	*Amount Given* (£)	0.28	0.79	0.57	1.05
		(0.46)	(0.99)	(1.57)	(1.31)
	Amount Given, % Exp. Fare	5	8	10	10
		(0.08)	(0.09)	(0.28)	(0.12)
	Number of Journeys	26	11	30	11
South Asian	*Amount Given* (£)	1.23	1.28	0.52	0.54
		(1.40)	(1.80)	(0.65)	(0.53)
	Amount Given, % Exp. Fare	23	12	9	5
		(0.26)	(0.17)	(0.12)	(0.05)
	Number of Journeys	9	11	14	7
All testers	*Amount Given* (£)	0.60	1.14	0.72	1.08
		(0.73)	(1.43)	(1.07)	(1.23)
	Amount Given, % Exp. Fare	11	11	13	10
		(0.13)	(0.15)	(0.19)	(0.12)
	Proportion of Journeys Completed	0.27	0.27	0.31	0.34
	Total number of journeys	95	48	93	47

Note. Standard deviations are in parentheses.

business mechanism, in models (4) and (8), we include an additional dummy variable, *BusC Given*, that takes a value of 1 if a business card was given and 0 otherwise. We include this dummy without any interactions in order to provide an easily interpretable estimate of its net effect on driver giving.

In each subsequent model, the number of explanatory variables is increased or varied in order to examine the robustness of the estimated treatment effects. The additional variables we use were those recorded by the testers, and outlined in Table 2, which we group into three distinct sets: field, city, and ride controls. The set of field controls includes the variable for traffic intensity (recorded on a 10-point scale: 1 if not busy, 10 if very busy) and a dummy controlling for the weather conditions (1 if raining). The set of city controls includes three dummies for the location where the journey was taken: Birmingham, Trafford, or Salford (1 if yes), with those in Manchester taken as the baseline. The set of ride controls includes dummies controlling for whether the driver offered to take the passenger to a cashpoint/ATM (1 if offered) and if he or she tried to engage in a conversation (1 if yes).

In addition, we examine driver-giving conditional on the testers' ethnicity. We include dummy variables that take a value of 1 (and 0 otherwise) if the tester was *Black*, *South Asian*, and/or *Male*. We also include three additional sets of control variables: driver, tester, and appearance controls. The set of driver and tester controls includes the drivers' and testers' gender and age. The set of appearance controls includes the

Figure 1. (Color online) Distribution of Giving, by Treatment

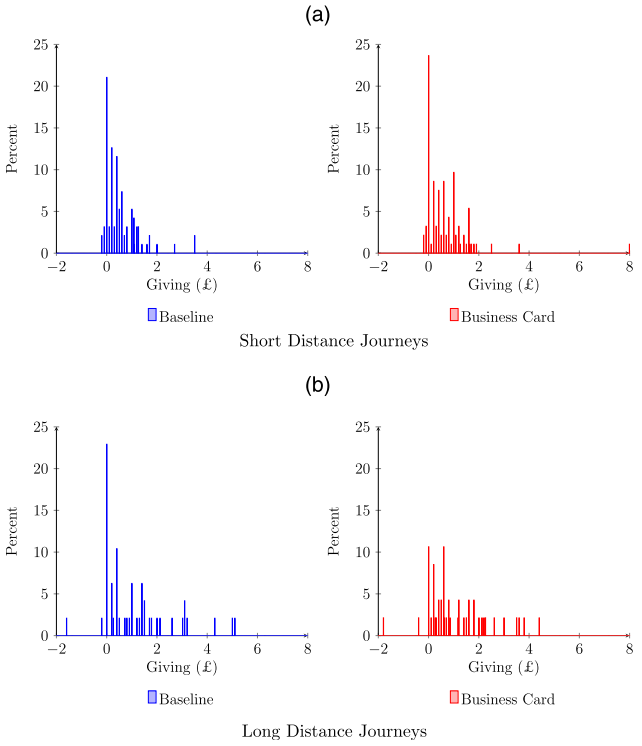


Table 5. Determinants of Driver Giving

	Random effects Tobit regressions											
	Amount given (in £) in models (1)–(4)				Amount given (% of the expected fare) in models (5)–(12)							
Dep. variable:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Long	0.676***	0.729***	0.734***		0.049	0.070	0.070		0.028	0.013	0.000	0.022
	(0.254)	(0.278)	(0.272)		(0.068)	(0.075)	(0.074)		(0.055)	(0.055)	(0.067)	(0.069)
BusC	0.072	0.042	0.048		0.038	0.030	0.033		0.024	0.031	0.034	0.027
	(0.157)	(0.159)	(0.156)		(0.042)	(0.042)	(0.042)		(0.043)	(0.043)	(0.043)	(0.043)
BusC × *Long*	0.034	0.062	0.034		0.011	0.026	0.023		0.023	0.013	0.005	0.034
	(0.295)	(0.301)	(0.295)		(0.078)	(0.080)	(0.080)		(0.077)	(0.076)	(0.076)	(0.078)
BusC Given				0.217				0.112**				
				(0.168)				(0.044)				
Black									–0.121***	–0.123***	–0.111**	–0.127**
									(0.046)	(0.046)	(0.046)	(0.050)
South Asian									–0.054	–0.059	–0.052	0.034
									(0.056)	(0.056)	(0.056)	(0.066)
Male									–0.097**	–0.095**	–0.087**	–0.135**
									(0.042)	(0.042)	(0.043)	(0.063)
Constant	1.243***	1.495***	1.388***	1.250***	0.293***	0.335***	0.310***	0.276**	0.286	0.266	0.357*	0.630
	(0.389)	(0.389)	(0.060)	(0.448)	(0.100)	(0.113)	(0.113)	(0.113)	(0.194)	(0.193)	(0.199)	(0.731)
Observations	283	282	281	280	283	282	281	280	275	274	274	274
Controls												
City	✓	✓	✓	✓	✓	✓	✓	✓			✓	✓
Field		✓	✓	✓		✓	✓	✓			✓	✓
Ride			✓	✓			✓	✓		✓	✓	✓
Driver									✓	✓	✓	✓
Tester									✓	✓	✓	✓
Appearance												✓

Notes. Standard errors are in parentheses. The number of observations falls slightly as more controls are included as a result of missing entries. Models (1)–(4) are left-censored at 0 and right-censored at the difference between the expected fare had the driver completed the journey and the amount paid by the tester. Models (1)–(8) include tester fixed effects. Models (5)–(12) are left-censored at 0 and right-censored at 1. All reported explanatory variables are dummy variables that take values of 1 in the following cases (and 0 otherwise): *Long*, if a long-distance journey; *BusC*, if the business card treatment; *BusC Given*, if a business card was given by the driver when asked; *Black*, if the tester is black; *South Asian*, if the tester is South Asian; and *Male*, if the tester is male. "Constant" denotes the estimate of the constant. "Obs." denotes the number of observations. None of the appearance controls are estimated to have a significant effect on giving at the 5% level.

***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

average rating the tester was given along each of the appearance dimensions we elicited, attractiveness, aggressiveness, friendliness, trustworthiness, and wealthiness, as reported in Table B6 in Online Appendix B.

Result 1. The majority of taxi drivers exhibit other-regarding preferences.

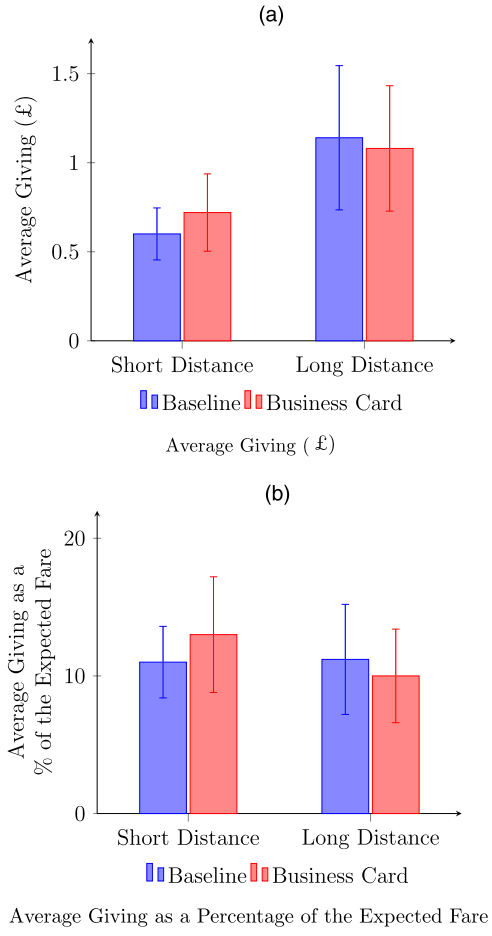
Support. Considering journeys from the baseline treatment, the null hypothesis of no giving can be rejected at the 1% level in both long- and short-distance journeys ($p < 0.01$, both cases, sign test). We find that 74% of drivers in these treatments give part of the journey for free, with 27% completing the journey in full. Similar findings are observed in the business card treatment, with 76% of drivers giving at least part of the journey for free, with 32% of all journeys being completed in full.

Result 2. Drivers' other-regarding preferences are well defined over relative payoffs.

Support. Examining journeys from the baseline treatment, average driver-giving is significantly different in short-distance journeys in comparison with long-distance journeys ($p = 0.079$, robust rank-order test). This is shown graphically in Figure 2(a). The distribution of giving is also found to vary by the distance of the journey ($p = 0.059$, Kruskal–Wallis test). Regressions (1)–(3) in Table 5 support these conclusions, reporting significant and positive coefficient estimates on the long-distance dummy ($p < 0.01$).

However, when giving as a percentage of the expected fare is considered, no significant differences are reported by distance ($p = 0.88$, robust rank-order test; see Figure 2(b)). Furthermore, the distance of the journey has no significant effect on the distribution of giving ($p = 0.86$, Kruskal–Wallis test), and no significant treatment effects are reported in models (5)–(12) when the dependent variable is giving as a percentage of the expected fare ($p > 0.1$ in all cases, in all

Figure 2. (Color online) Average Giving



Note. Vertical bars represent 95% confidence intervals.

regressions in Table 5). This suggests that giving, relative to the length of the journey, is constant.

Result 3. Asking for a business card has no effect on drivers' behavior. However, conditional on giving a business card, drivers give significantly more.

Support. Comparing average giving between business card and baseline treatments, no significant differences are reported between the short- and long-distance treatments, respectively ($p = 0.34$ and $p = 0.67$, robust rank-order tests). Similarly, asking for a business card has no significant impact on the distribution of giving in either short- or long-distance treatments, respectively ($p = 0.67$ and $p = 0.44$, Kruskal–Wallis test). The same is true for giving as a percentage of the expected fare, with no significant differences found between business card and baseline treatments in short- or long-distance journeys or when journeys are pooled ($p > 0.1$ in all cases, robust rank-order tests). Estimates from Table 5 support these results, with the coefficient on the *BusC* dummy found to be not significant at conventional levels across regressions ($p > 0.1$ in all cases). The estimates from Table 5 also outline how drivers who do give a business card do not give significantly more of the journey for free in absolute terms than those who do not (model (4), $p > 0.1$), but they do give more as a percentage of the expected fare (model (8), $p < 0.05$).[24]

Result 3 is supportive of the idea that drivers who ply for hire at taxi ranks treat the interactions they have with passengers as one shot and thus that our baseline interactions are not confounded with reputational concerns. Result 3 also suggests that repeated business is not a major concern for the drivers because repeated interaction effects are not triggered when they are asked for a business card. This is likely because the majority of drivers choose not to provide one, with only 45% deciding to provide one when asked (see Table 2). However, if the tester receives a business card, the driver does give significantly more.[25] This suggests repeated interaction effects could play a role in fostering other-regard in this setting but first need to be successfully "activated."[26]

We now consider the effect that the testers' ethnicity has on the drivers' behavior and conduct pairwise comparisons of drivers giving to black, white, and South Asian testers. In line with the previous literature, our primary concern is in considering differentials between black and white testers. In addition, as we have fewer South Asian observations in comparison with the number of black and white observations (41 journeys compared with 78 and 164 journeys, respectively), the results and discussions related to the South Asian testers should be interpreted with care.

To consider the effect of ethnicity on journey completions, Table 6 reports the estimated coefficients and marginal effects from a number of random effects probit regressions, where the dependent variable is a dummy that takes a value of 1 if the journey was completed and 0 otherwise. We increase the number of explanatory variables in each subsequent model and use the same sets of control variables as outlined in Table 5.

Result 4. Drivers' other-regard is smallest when the tester is black.

Support. Comparing the proportion of journeys completed by testers' ethnicity, we find that black testers have their journey completed significantly less often than white and South Asian testers, respectively, in the baseline treatment ($p = 0.045$ and $p = 0.088$, Fisher's exact test; see also Figure 3). No significant differences are reported between white and South Asian testers ($p = 0.793$, Fisher's exact test). The estimates of random effects probit regressions in Table 6 outline how the estimated coefficient on the black dummy is

Table 6. Determinants of Journey Completion

	Random effects probit regressions				
	Journey completed				
Dependent variable:	(1)	(2)	(3)	(4)	(5)
Black	–0.494**	–0.509**	–0.544**	–0.512**	–0.650***
	(0.198)	(0.217)	(0.221)	(0.223)	(0.246)
South Asian	–0.313	–0.419	–0.450*	–0.438	–0.124
	(0.242)	(0.267)	(0.268)	(0.271)	(0.326)
Male		–0.335*	–0.335*	–0.326	–0.344
		(0.190)	(0.193)	(0.201)	(0.310)
Constant	–0.913	0.152	0.074	0.293	3.658
	(0.760)	(0.862)	(0.877)	(0.919)	(3.621)
Observations	274	274	273	273	273
Marginal effects					
Black	–0.160***	–0.161**	–0.169***	–0.159**	–0.188***
	(0.060)	(0.065)	(0.064)	(0.065)	(0.064)
South Asian	–0.107	–0.136*	–0.143*	–0.139*	–0.041
	(0.078)	(0.081)	(0.079)	(0.079)	(0.106)
Male		–0.120*	–0.118*	–0.114*	–0.116
		(0.066)	(0.066)	(0.069)	(0.104)
Controls					
Treatment	✓	✓	✓	✓	✓
Driver	✓	✓	✓	✓	✓
Tester		✓	✓	✓	✓
Ride			✓	✓	✓
Field				✓	✓
City				✓	✓
Appearance					✓

Notes. Standard errors are in parentheses. Marginal effects are evaluated where the passengers' ethnicity is white, with all other variables evaluated at the mean. All reported explanatory variables are dummy variables that take values of 1 in the following cases (and 0 otherwise): *Black*, if the tester is black; *South Asian*, if the tester is South Asian; and *Male*, if the tester is male. *Treatment* controls includes a dummy for long-distance journeys, a dummy for the business card treatment, and a dummy for whether a business card was received, along with interactions. Estimates from the *Appearance* controls are not reported for space concerns and because none are found to be significant at the 5% level.

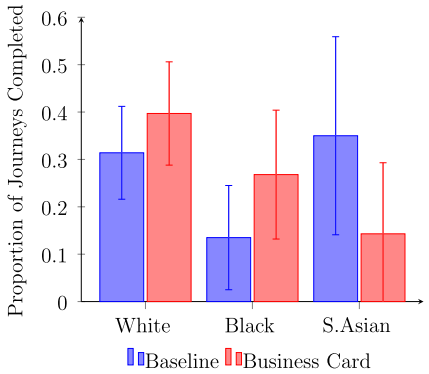
***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

negative and significant ($p < 0.01$). This estimate is robust to specification changes and becomes increasingly significant as more controls are included. The estimated marginal effect size is robust across models and is estimated to be highly significant ($p < 0.01$ in all cases). Similar to Table 5, none of the appearance characteristics is found to be significant at the 5% level.

Pairwise comparisons of average giving by drivers to white, black, and South Asian testers in the baseline treatment reveals no significant differences between white and South Asian testers in the short- or long-distance treatments, respectively ($p = 0.41$ and $p = 0.88$, robust rank-order tests). However, significant differences between white and black testers are reported in the short- but not in the long-distance treatment, respectively ($p = 0.001$ and $p = 0.37$, robust rank-order tests). Similarly, a significant difference between South Asian and black testers is found in the short- but not in the long-distance treatment, respectively ($p = 0.07$ and $p = 0.47$, robust rank-order tests).

Considering the amount given as a percentage of the expected fare reveals that both white and South Asian testers are given significantly more than black testers, respectively ($p = 0.002$ and $p = 0.045$, robust rank-order tests), but no differences are found between

Figure 3. (Color online) Proportion of Journeys Completed, by Tester Ethnicity



Note. Vertical bars represent 95% confidence intervals.

white and South Asian testers, respectively ($p = 0.56$, robust rank-order test). The estimates in Table 5 further support the nonparametric results: across regressions (5)–(12), the coefficient on the black dummy is negative, highly significant ($p < 0.05$ in all cases, Wald tests), and robust to changes in the model specification. None of the appearance controls included in model (12) in Table 5 is found to have a significant effect on giving.

The differential treatment of testers by ethnicity remains in the business card treatment, with white testers receiving more than black testers in the short-distance treatment ($p < 0.001$, robust rank-order test), although no difference is observed between white and South Asian testers ($p = 0.63$, robust rank-order tests). No differences are reported between black and South Asian testers in either distance treatment ($p > 0.1$, robust rank-order test in both cases). Comparing giving as a percentage of the expected fare reveals differences in giving between white and black and white and South Asian testers, respectively ($p < 0.001$ and $p = 0.02$, robust rank-order tests) but no difference between black and South Asian testers, respectively ($p = 0.817$, robust rank-order test).[27]

Result 4 outlines how black testers are treated significantly worse than white and South Asian testers. Indeed, the inclusion of additional controls, in particular, the appearance characteristics, increases the magnitude of the coefficient of the black dummy in both the tobit and probit regressions. Furthermore, despite the significant differences in appearance characteristics between ethnicities, they are estimated to have no significant effect on journey completion. Despite their subjectivity, the use of these characteristics serves to demonstrate that the results are robust even when controlling for significant differences in appearance characteristics between social groups. The evidence is consistent with taste-based discrimination manifested in social preferences, although, as with any observed disparity between social groups in a strategic setting, they can also potentially be rationalized as statistical discrimination.[28]

4.3. Structural Models

The reduced-form estimates given in Section 4.2 provide evidence of variation in driver giving that is conditional on the testers' ethnicity. However, they do not provide quantitative estimates of the preferences that we assume to underpin this behavior. We estimate the parameters of a number of utility functions in order to link our empirical analysis to behavioral theory that assumes that individuals differentiate between social groups as a consequence of taste-based discrimination. Details of the estimation strategy and the estimates themselves are given in Online Appendix B.

The structural estimates, given in Table B2 in Online Appendix B, confirm the reduced-form estimation results and suggest that other-regarding preferences are group contingent. Because the taxi drivers face a situation that is analogous to laboratory dictator games, we use the preference estimates to predict the drivers' behavior in a hypothetical £10 dictator game. These predictions are given in Table B3 in Online Appendix B. By doing this, we can benchmark the predictions our estimates produce against the behavior observed in laboratory experiments in the literature. We find that our estimates produce behavior that is consistent with the other-regarding preference literature.

5. Discussion

Although the results in Section 4 are assumed to be a consequence of drivers having other-regarding preferences, there are other potential explanations. The extent of giving can be explained by (1) stopping distances associated with finding somewhere safe to drop the passenger, (2) convenience for the driver associated with considerations regarding his or her next journey, (3) the drivers' experience and expectations of passengers bargaining, and (4) social pressure. In addition, the observations relating to ethnicity could be an artifact of multiple hypothesis testing. Although (1) and (2) may seem similar, we feel it is important to separate the transactional explanation of the drivers' behavior, explanation (1), from the strategic motivation, explanation (2), that could stem from expectations about the next customer. We examine each of these alternative explanations individually and find little evidence for their support. Even when taking the evidence jointly, the main conclusion of our paper remains unchanged.

To examine (1), we begin by making the assumption that *some* giving may be an artifact of the drivers finding a convenient location for the passenger to alight. We do this by assuming that the driver requires either one, two, or three additional charges on the meter (approximately 190, 380, and 570 yards, respectively) in order to find a suitable location.[29] For example, a driver recorded as giving £0.20 may actually have given nothing but required £0.20 worth of metered distance (approximately 190 yards or a single charge on the meter) in order to find a location to stop. Thus, as a robustness check, giving is redefined as the amount of the journey that is given for free *minus* the amount the driver is induced to give as a consequence of the assumed stopping distance. This is in addition to the driver already having £1 worth of metered distance as a notice period in the short-distance treatment and £2 worth of metered distance as a notice period in the long-distance treatment in order

Table 7. Stopping Distances and Driver Giving

	Amount given (£)		
Assumed stopping distance	Short	Long	Percent giving more than £0
190 yards	0.459***	1.028***	71.3
	(0.68)	(1.329)	
380 yards	0.347***	0.896***	58.0
	(0.626)	(1.275)	
570 yards	0.266***	0.787***	42.7
	(0.564)	(1.208)	

Notes. Standard deviations are in parentheses. Amounts given only include baseline observations. When giving defined in this way produces a negative amount given, we assume that the driver decided to give nothing.

*** denotes significance at the 1% level.

to stop the taxi.[30] Table 7 presents average giving under the three different assumptions about stopping distance using observations from the baseline, short-, and long-distance treatments.

As can be seen in Table 7, giving is still significantly different from zero for both short- and long-distance journeys regardless of the assumed stopping distance. Drivers also give significantly more in the long-distance treatment in comparison with the short-distance treatment under all three stopping distance assumptions ($p = 0.036$, $p = 0.037$, and $p = 0.037$, robust rank-order tests). Furthermore, the majority of drivers still give more than zero, except when a conservative stopping distance of 570 yards is assumed. Even then, almost half of all drivers still give positive amounts. As such, we conclude that our results are robust to stopping distance confounds.

To shed light on (2) and (3), we conducted a survey of 50 taxi drivers from ranks used within the study and 65 passengers that were queuing for a taxi, and we observed the behavior of 97 passengers entering taxis from a rank.[31] To understand the drivers' routine, drivers were asked to indicate the number of daily journeys, how many of these journeys start at a taxi rank, and what they believe the average fare is. Drivers were also asked about the expected fare of a sample short- and long-distance journey, where the sample journeys were journeys that we used within the study.[32]

To address (2), drivers were asked a multiple-choice question about what they did on completing a journey: return to a home rank, return to a different rank, cruise and look for a passenger, or do something else. This question was designed to shed light on the drivers' opportunity cost of giving. Because all journeys that were taken in the experiment involved driving away from a home rank, returning to the rank would have been more costly, in terms of both time and fuel, if the driver had continued to drive past the amount the passenger could afford.

To address (3), drivers were asked if they would be willing to bargain over the journey specified *before* the journey began and the lowest fare they would accept if they were willing to negotiate. In addition, they were asked if they would be willing to bargain with a passenger who was inside the taxi. Passengers were asked if they ever bargained with the driver when catching a taxi from the rank.

The drivers' responses are presented in Table 8, panel A. The passengers' responses and the recorded observation results are presented in panel B. The responses in Table 8 highlight three main points relating to (2). First, the vast majority of taxi journeys are taken from ranks (86%), indicating a low proportion being hailed. Second, the majority of drivers report returning to a home rank (74%). This suggests that giving to the passenger, by continuing to drive away from the rank, is not done at the drivers' convenience. On the contrary, driving away from the rank is analogous to driving away from the next passenger, and therefore, it is costly.

Addressing (3), we note from Table 8 that the vast majority of drivers (96% for a hypothetical short-distance journey and 88% for a hypothetical long-distance journey) said they would not bargain with a passenger before the passenger was inside the vehicle. The lowest fare they would accept is also above the amount our testers could afford. In addition, the majority of drivers would refuse to bargain with passengers midjourney (96%). Drivers' expected fare estimates are also in line with our own calculations, suggesting that they can accurately calculate how much each journey will cost. This is perhaps unsurprising because the "knowledge" tests that drivers operating in Britain are required to take (and pass) have been shown to change the structure of their brains (Maguire et al. 2000), allowing them to more easily map and plan a route. Our survey and observation of passengers also show that the desire to negotiate is limited, with only a single passenger observed attempting to bargain with a driver and

Table 8. Driver and Passenger Survey Responses

Panel A: Driver survey, $N = 50$		
Total	*No. of daily journeys*	12.94 (4.47)
	No. of journeys that start at a rank	11.26 (4.45)
	Average fare (£)	6.40 (1.38)
	Percent of journeys that start at a rank[a]	86.14 (0.12)
	Modal taxi model	LTI TXII
Short-distance journeys	*Expected fare* (£)	6.17 (0.78)
	Willing to bargain? (1 if yes)	0.06 (0.24)
	Lowest fare if willing (£)	4.73 (2.11)
	Willing to bargain inside the taxi? (1 if yes)	0.04 (0.20)
	On completion	
	Return to home rank (%)	74
	Return to a different rank (%)[b]	16
	Cruise (%)	10
Long-distance journeys	*Expected fare* (£)	11.85 (1.97)
	Willing to bargain? (1 if yes)	0.12 (0.33)
	Lowest fare if willing (£)	9.33 (1.03)
	Willing to bargain inside the taxi? (1 if yes)	0.04 (0.20)
	On completion	
	Return to home rank (%)	76
	Return to a different rank (%)	10
	Cruise (%)	10
Panel B: Passenger survey		
Do you bargain? (1 if yes), $N = 65$		0.03 (0.181)
Observed bargaining (1 if yes), $N = 97$		0.01 (0.1)

Notes. All responses relate to journeys taken between 9 a.m. and 5 p.m. Standard deviations are in parentheses.

[a]This percentage is calculated from the number of journeys that start at a rank and the number of journeys taken in day, and it was not a question on the questionnaire.

[b]The majority of drivers specifying this response outlined that they would return to different rank in the center of the city.

only two reporting that they have ever bargained with drivers over fares. Therefore, it seems unlikely that drivers are accustomed to bargaining with passengers because the vast majority of journeys are not bargained over.[33]

Interpretation (4) implies that drivers are concerned about behaving unkindly or dislike confrontation, and they therefore give despite having a preference not to. This would resonate with the conclusion of DellaVigna et al. (2012). However, there are a number of potential issues with this interpretation. First, *not* giving away goods and services for free in a market setting is unlikely to be perceived as unkind. For example, in an analogous situation in a retail setting, it is unlikely that a customer who wished to purchase 10 units of an item but who could only afford to buy 8 would regard a seller as unkind if the seller refused to give the customer 2 additional units for free. This contrasts with charitable giving, where giving to those who need it might be viewed as a normative action. Second, about 70% of drivers do not complete the journey but choose to eject the passenger before reaching the destination, and it seems unlikely that differences in perceived social pressure can explain the within-treatment variation in giving, given the standardization of the interactions, or can explain why giving is found to be proportional to the length of the journey. Third, in the context of our study, passengers could easily have taken an alternative and cheaper mode of transport, or they could have walked the final portion of the journey they could not afford.

Finally, to address concerns related to multiple hypothesis testing, we adjust the calculated p-values using the Holm–Bonferroni procedure (Holm 1979). These corrections, along with a description of the procedure, are presented in Section B.6 of Online Appendix B. We find that when correcting the p-values for the parametric analysis, Results 1–4 are found to be robust. Furthermore, when correcting the p-values from the nonparametric tests for Result 4, applying the most conservative correction procedure possible, we find that the black/white disparity in giving remains significant. However, considering just the nonparametric statistics, we cannot distinguish any of the comparisons that involve South Asian testers from type I error.

6. Conclusion

We report evidence that the majority of taxi drivers behave in an other-regarding manner in a market setting with limited possibilities for repeated interactions. Through our experimental design, we find evidence that taxi drivers have well-defined preferences over relative payoffs, a finding that resonates with the results of numerous laboratory experiments and behavioral theories of social preferences. We show that our findings are robust to a wide range of controls and a variety of potential behavioral and statistical confounds.

Variation in the ethnicity of the driver and the tester also allows us to explore recent theories of taste-based discrimination—namely, that other-regarding preferences are group contingent. We find evidence that the drivers' propensity to give is significantly smaller

when the passenger is black. This result is robust to controlling for variation in the testers' appearance, variation that otherwise could be driving the result. It is also robust to correcting for potential multiple comparison problems.

Although our results resonate with the group-contingent social preference hypothesis, as with other studies on discrimination, we acknowledge that we cannot rule out statistical motivations. However, our study distinguishes itself from previous work that examines discriminatory behavior in an important way. Whereas previous observations of differential treatment can be entirely explained by statistical motivations that are consistent with profit maximization absent social preferences, the behavior observed in our field experiment necessitates social preferences even if some sort of statistical discrimination exists. This is because helping those one deems deserving of help requires social preferences as well as some form of discrimination.

We acknowledge that markets where transactions are automated or done through a computer, such as asset and financial markets are unlikely to see the types of behaviors observed here. This is because the nature of the interaction between buyer and seller does not allow for such preferences to be expressed because market agents are not given the opportunity to behave in such a manner. However, many other types of markets exist. Especially in markets where bilateral face-to-face interactions are commonplace, we expect other-regarding preferences to play a much greater role than previously suggested. Other-regarding preferences may persist despite regulation and could possibly only be overcome by centralization.

Acknowledgments

The authors are grateful to seminar participants at the University of Auckland, Bielefeld University, University of Exeter, Humboldt University, University of Illinois Urbana–Champaign, University of Innsbruck, University of Michigan, University of Nuremberg, Universitat Pompeu Fabra, Royal Holloway University of London, Texas A&M University, Waseda University, attendees at the Economic Science Association of North American Meetings in Dallas, the Advances in Field Experiments Conference at the University of Chicago, the 2016 Asia Meeting of the Econometric Society in Kyoto, and the 2016 Royal Economic Society Meetings for helpful comments and suggestions. The authors also thank Loukas Balafoutas, David Reiley, and Henry Schneider for comments.

Endnotes

[1] See Guala and Filippin (2017) and Zizzo (2010, 2012) for some critiques of the study of discrimination in the laboratory.

[2] Under the taxonomy of Harrison and List (2004), our experiment is classified as a natural field experiment.

[3] The taxi markets we study satisfy all the requirements of a marketplace, as discussed by Al-Ubaydli and List (2019).

[4] This is the case in the cities that we study, but it varies throughout Great Britain.

[5] Local authorities use the licensing of Hackney carriage drivers in order to restrict the entry of new drivers. This is done in an effort to stop supply exceeding demand.

[6] Tipping is not expected in Great Britain, as it is in other countries. In a taxi, it is often common to round up to the nearest pound to facilitate the payment.

[7] Many drivers are, however, affiliated with a firm from which they can take private hire bookings.

[8] This figure is taken from the 2015 Unmet Taxi Demand Survey published by the Manchester City Council.

[9] In line with the ethics guidelines at the University of Exeter, we invited everyone who applied for the job to an interview and made job offers to everyone who attended the interview. This procedure differs from that of Ayres and Siegelman (1995), where the experimenters selected the testers based on their own perception of "average attractiveness."

[10] For example, Balafoutas et al. (2013) employed five testers.

[11] We discussed the short-distance/baseline treatment, which is described in Section 3.2.

[12] Once the study was completed, all the testers were asked to guess what they thought the study was about. None correctly identified the research questions.

[13] As such, we follow the recruitment procedure used by Landry et al. (2006) closely.

[14] The photo rating sessions were conducted at the end of other, unrelated experimental sessions conducted at the University of Exeter.

[15] Table B5 in Online Appendix B presents the correlations between the testers' perceived facial appearance characteristics.

[16] In UK taxis, all passengers are required to sit in the back.

[17] For example, a tester would be required to take a journey from rank *A* to location *X* and then walk (up to 15 minutes) to another rank in order to take the next journey. We omit the routes to ensure the anonymity of the drivers. For a sample of destinations, see Online Appendix A.

[18] Some of the destinations in the long-distance treatment did have taxi ranks close by. However, for these journeys, because the drivers were traveling between "local authorities," they were prohibited from picking up passengers at the destination.

[19] The first ride taken by each tester was discreetly observed by the experimenter to ensure that he or she entered the taxi correctly.

[20] We did not use recording devices to monitor testers because of the simple, highly stylizsed nature of the interaction. This is in line with procedures used in other prominent studies that did not use recording devices, such as Ayres and Siegelman (1995), Landry et al. (2006), and Balafoutas et al. (2013). Castillo et al. (2013) study a complex bargaining interaction and, as far as we are aware, is the only study that uses recording devices to monitor confederates. Although important for their study, we felt the ethical considerations outweighed any potential benefits in our case. A consequence of this is that the testers were not fully monitored and only had to hand in the receipts for the journeys they took.

[21] This cannot be inferred from the receipts, which only contain information about the amount paid by the tester.

[22] Although the Manchester Council does not collect driver ethnic demographics, the Birmingham City Council provided the following information regarding the distribution of driver ethnicities (obtained from a Freedom of Information request, number FOI 15327): 82.6% South Asian, 9.6% white, 3.8% black, and 3.9% other. "Other" includes all drivers who declared other ethnicities (e.g., mixed) and

those who did not disclose their ethnicity. The ethnic distribution of our sample is representative of the population distribution.

[23] It is worth pointing out what the experimental procedure was *not*. The procedure was not an attempt to obtain free journeys by demanding them from the driver, nor did the testers maneuver the driver into making a decision he or she did not want to make. The testers were instructed to respect the driver at all times, and at no point did the testers question the drivers' right to charge the metered fare. As the tester requests the reduction of the fare, the driver clearly possesses the right to grant or refuse the request and charge the metered amount: the interaction cannot be interpreted as a negotiation.

[24] One could argue that there exists a selection effect in the business card treatment. Those who are asked for a business card, and decide to give one, are potentially different from those that are asked and do not give a business card. Because drivers are randomly assigned to treatments, these types should also be present in the baseline treatment. However, because both being asked for and giving a business card seem to be required to trigger reputational concerns, we can rule out reputational concerns as a confound in the Baseline treatment. We thank an anonymous referee for pointing this out.

[25] Giving a business card is not found to be significantly correlated with any of the elicited appearance characteristics or the gender of the testers, the demographics of the driver, or the ride characteristics.

[26] In addition, we find some evidence that a driver who tries to start a conversation with the passenger (even though the passenger did not engage) does give significantly more of the journey for free. Other control variables, such as traffic intensity and weather, are not found to be correlated with giving.

[27] To test to see whether our results are mainly due to appearance differences or driven by particular individuals, we compare each of the 12 white testers with each of the 7 black testers and note whether the white testers get more than the black testers (in terms of both pounds and the amount given as a percentage of the expected fare). We find that when comparing giving in pounds, in 69 of 84 comparisons, the white tester is given more. When comparing giving as a percentage of the expected fare, the white tester is given more in 71 of the 84 comparisons. These comparisons are given in Tables B10 and B11 of Online Appendix B.

[28] For instance, drivers might want to help those in need regardless of race but might have different beliefs about the need of the passenger. They might think that some groups are more honest than others or simply have coarser beliefs for those populations with which they interact less frequently.

[29] See Table A1 of Online Appendix A for the exact amounts.

[30] This is because drivers are told about the amount the tester can afford when the meter reaches £3 in the short-distance treatment and £6 in the long-distance treatment.

[31] The survey and observations were conducted in Manchester. The questionnaire is given in Online Appendix A.

[32] Drivers were also asked to report their income, but the majority refused to disclose this information.

[33] Passengers are allowed to bargain with drivers ex ante or before they enter the taxi. However, if no discount is agreed prior to the journey beginning, the driver is allowed to charge the metered fare by law.

References

Al-Ubaydli O, List JA (2019) How natural field experiments have enhanced our understanding of unemployment. *Nat. Hum. Behav.* 3:33–39.

Ayres I, Siegelman P (1995) Race and gender discrimination in bargaining for a new car. *Amer. Econom. Rev.* 85(3):304–321.

Balafoutas L, Beck A, Kerschbamer R, Sutter M (2013) What drives taxi drivers? A field experiment on fraud in a market for credence goods. *Rev. Econom. Stud.* 80(3):876–891.

Bandiera O, Barankay I, Rasul I (2005) Social preferences and the response to incentives: Evidence from personnel data. *Quart. J. Econom.* 120(3):917–962.

Becker GS (1971) *The Economics of Discrimination*, 2nd ed. (University of Chicago Press, Chicago).

Benson PL, Karabenick SA, Lerner RM (1976) Pretty pleases: The effects of physical attractiveness, race, and sex on receiving help. *J. Experiment. Soc. Psych.* 12(5):409–415.

Benz M, Meier S (2008) Do people behave in experiments as in the field? Evidence from donations. *Experiment. Econom.* 11(3): 268–281.

Bertrand M, Mullainathan S (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Amer. Econom. Rev.* 94(4):991–1013.

Camerer C, Babcock L, Loewenstein G, Thaler R (1997) Labor supply of New York City cabdrivers: One day at a time. *Quart. J. Econom.* 112(2):407–441.

Camerer C, Fehr E (2004) Measuring social norms and preferences using experimental games: A guide for social scientists. Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, Gintis H, eds. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies* (Oxford University Press, Oxford, UK), 55–95.

Camerer C, Babcock L, Loewenstein G, Thaler R (1997) Labor supply of New York City cabdrivers: One day at a time. *Quart. J. Econom.* 112(2):407–441.

Castillo M, Petrie R, Torero M, Vesterlund L (2013) Gender differences in bargaining outcomes: A field experiment on discrimination. *J. Public Econom.* 99(March):35–48.

Chen R, Chen Y (2011) The potential of social identity for equilibrium selection. *Amer. Econom. Rev.* 101(6):2562–2589.

Chen Y, Li SX (2009) Group identity and social preferences. *Amer. Econom. Rev.* 99(1):431–457.

Cooper D, Kagel JH (2009) Other-regarding preferences: A selective survey of experimental results. Kagel JH, Roth AE, eds. *Handbook of Experimental Economics*, vol. 2 (Princeton University Press, Princeton, NJ), 217–289.

DellaVigna S, List JA, Malmendier U (2012) Testing for altruism and social pressure in charitable giving. *Quart. J. Econom.* 127(1):1–56.

Drouvelis M, Nosenzo D (2013) Group identity and leading-by-example. *J. Econom. Psych.* 39(December):414–425.

Dufwenberg M, Heidhues P, Kirchsteiger G, Riedel F, Sobel J (2011) Other-regarding preferences in general equilibrium. *Rev. Econom. Stud.* 78(2):613–639.

Gneezy U, List J, Price MK (2015) Toward an understanding of why people discriminate: Evidence from a series of natural field experiments. NBER Working Paper No. 17855, National Bureau of Economic Research, Cambridge, MA.

Goette L, Huffman D, Meier S (2006) The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *Amer. Econom. Rev.* 96(2):212–216.

Guala F, Filippin A (2017) The effect of group identity on distributive choice: Social preference or heuristic? *Econom. J.*127(602): 1047–1068.

Harrison GW, List JA (2004) Field experiments. *J. Econom. Lit.* 42(4): 1009–1055.

Heckman JJ (1998) Detecting discrimination. *J. Econom. Perspect.* 12(2): 101–116.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.* 6(2):65–70.

Landry CE, Lange A, List JA, Price MK, Rupp NG (2006) Toward an understanding of the economics of charity: Evidence from a field experiment. *Quart. J. Econom.* 121(2):747–782.

Levitt SD (2004) Testing theories of discrimination: Evidence from the weakest link. *J. Law Econom.* 47(2):431–452.

List JA (2004) The nature and extent of discrimination in the marketplace: Evidence from the field. *Quart. J. Econom.* 119(1): 49–89.

List JA (2006) The behavioralist meets the market: Measuring social preferences and reputation effects in actual transactions. *J. Political Econom.* 114(1):1–37.

Maguire EA, Gadian DG, Johnsrude IS, Good CD, Ashburner J, Frackowiak RS, Frith CD (2000) Navigation-related structural change in the hippocampi of taxi drivers. *Proc. Natl. Acad. Sci. USA* 97(8):4398–4403.

Mitra A, Ray D (2014) Implications of economic theory of conflict: Hindu–Muslim violence in India. *J. Political Econom.* 122(4):719–765.

Mobius MM, Rosenblat TS (2006) Why beauty matters. *Amer. Econom. Rev.* 96(1):222–235.

Mujcic R, Frijters P (2013) Still not allowed on the bus: It matters if you're black or white! IZA Discussion Paper 7300, IZA, Bonn, Germany.

Öhman A (1986) Face the beast and fear the face: Animal and social fears as prototypes for evolutionary analyses of emotion. *Psychophysiology* 23(2):123–145.

Phelps ES (1972) The statistical theory of racism and sexism. *Amer. Econom. Rev.* 62(4):659–661.

Schneider HS (2012) Agency problems and reputation in expert services: Evidence from auto repair. *J. Indust. Econom.* 60(3):406–433.

Schupp HT, Öhman A, Junghöfer M, Weike AI, Stockburger J, Hamm AO (2004) The facilitated processing of threatening faces: An ERP analysis. *Emotion* 4(2):189–200.

Stoop J (2014) From the laboratory to the field: Envelopes, dictators and manners. *Experiment. Econom.* 17(2):304–313.

Stoop J, Noussair CN, Van Soest D (2012) From the laboratory to the field: Cooperation among fishermen. *J. Political Econom.* 120(6): 1027–1056.

Tajfel H, Billig MG, Bundy RP, Flament C (1971) Social categorization and intergroup behaviour. *Eur. J. Soc. Psych.* 1(2):149–178.

van Der Mewe GW, Burns J (2008) What's in a name? Racial identity and altruism in post-apartheid South Africa. *South African J. Econom.* 76(2):266–275.

Winking J, Mizer N (2013) Natural-field dictator game shows no altruistic giving. *Evolution Human Behav.* 34(4):288–293.

Xiao E, Houser D (2005) Emotion expression in human punishment behavior. *Proc. Natl. Acad. Sci. USA* 102(20):7398–7401.

Zizzo DJ (2010) Experimenter demand effects in economic experiments. *Experiment. Econom.* 13(1):75–98.

Zizzo DJ (2012) Inducing natural group identity: A RDP analysis. University of East Anglia Discussion Paper 12-03, University of East Anglia, Norwich, UK.